Measuring the Power of Learning.®

Research Report
ETS RR–17-12

Performance of Automated Speech Scoring on Different Low- to Medium-Entropy Item Types for Low-Proficiency English Learners

Anastassia Loukina

Klaus Zechner

Su-Youn Yoon

Mo Zhang

Jidong Tao

Xinhao Wang

Chong Min Lee

Matthew Mulholland

March 2017

ETS Research Report Series

RESEARCH REPORT

Performance of Automated Speech Scoring on Different Low- to Medium-Entropy Item Types for Low-Proficiency English Learners

Anastassia Loukina, Klaus Zechner, Su-Youn Yoon, Mo Zhang, Jidong Tao, Xinhao Wang, Chong Min Lee, & Matthew Mulholland

Educational Testing Service, Princeton, NJ

This report presents an overview of the *SpeechRater*^{SM} automated scoring engine model building and evaluation process for several item types with a focus on a low-English-proficiency test-taker population. We discuss each stage of speech scoring, including automatic speech recognition, filtering models for nonscorable responses, and scoring model building and evaluation and compare how the performance at each step differs between different item types. We conclude by discussing the effect of item type on automated scoring performance. We also give recommendations about what considerations should be taken into account when developing tests for low-proficiency English speakers to obtain reliable scores from an automatic scoring engine.

Keywords Automated speech scoring; constructed response scoring; *SpeechRater*^{SM}; filtering model

doi:10.1002/ets2.12139

Automated scoring of non-native speech dates back to the early 1990s, at which time most systems focused only on very limited aspects of speech, for example, pronunciation or fluency (Bernstein, Cheng, Suzuki, Ave, & Alto, 2010; Cucchiarini, Strik, & Boves, 1997; Franco, Neumeyer, Digalakis, & Ronen, 2000). Additionally, owing to limitations of automatic speech recognition (ASR) systems in those days, the focus was on restricted or predictable speech, such as reading sentences aloud or repeating short sentences aloud, presented acoustically (Bernstein, De Jong, Pisoni, & Townshend, 2000; Townshend, Bernstein, Todic, & Warren, 1998).

Since the early 2000s, several groups have built systems for scoring less constrained and more unpredictable speaking items, which incorporated additional sources of information for scoring, for example, diversity of vocabulary or grammatical complexity (Bernstein et al., 2010; Chen & Zechner, 2011; Strik, Van De Loo, Van Doremalen, & Cucchiarini, 2010; Yoon, Bhat, & Zechner, 2012; Zechner, Higgins, Xi, & Williamson, 2009). Recent work has also looked at evaluating the content relevance of spoken responses (Loukina, Zechner, & Chen, 2014; Somasundaran, Lee, Chodorow, & Wang, 2015; Xie, Evanini, & Zechner, 2012).

There have been few studies in the past on how an automated speech scoring system performs differentially across a set of diverse item types, ranging from highly predictable to relatively open-ended speech elicited from a test taker, with few exceptions described in the following paragraphs.

Cheng, D'Antilio, Chen, and Bernstein (2014) presented an automated speech scoring system for an Arizona K–12 language test for English-language learners that contained a range of items, from predictable (read or repeated prompt) to more open ended. There were 11 different item types, of which 9 were open ended. The automated scoring system used features related to fluency, pronunciation accuracy, vocabulary, and content. Item type-level human–machine correlations varied from .6 to .9 depending on both item type and grade level. They reported that for most open-ended items, machine performance was at a level similar to or better than human scoring. They also found that machine performance was significantly lower than human–human agreement for short-response item types and the items that requested the test taker to repeat the prompt.

Zechner et al. (2014) described a speech scoring system combining machine and human scores in a hybrid approach to score spoken English in an achievement test (unlike a more general proficiency test, an achievement test measures the test

Corresponding author: A. Loukina, E-mail: aloukina@ets.org

taker's knowledge based on a previously studied curriculum) for international teachers of English as a foreign language. The eight item types included in this study ranged from very predictable (e.g., read aloud) to moderately predictable (e.g., produce a sentence using some keywords). The correlation coefficients between machine and human scores ranged from .3 to .7 by item type, and there was no consistent pattern of difference between low- and medium-entropy items. They also reported the word error rate (WER) for the speech recognizer: WER was lower for predictable items (around 10%) than for moderately predictable items (30–35%).

Finally, Evanini, Heilman, Wang, and Blanchard (2015) reported on applying existing automated speech scoring technology to the three item types of the speaking section of the *TOEFL Junior*® test's comprehensive assessment, targeting students in the range of about 11–17 years of age. Item types comprise both predictable (read aloud) and more open-ended speech (e.g., picture narration). The system performance ranged from .6 to .7 for machine–human score correlation coefficients for the three item types examined in the study.

Although these three studies differ in terms of item type, test products, and test-taker populations, their results show some similarities: In both Zechner et al. (2014) and Evanini et al. (2015), the WER of the speech recognizer was lower for predictable items than for moderately predictable items, as one may also expect from general work on speech recognition. Furthermore, in both Cheng et al. (2014) and Zechner et al. (2014), sentence-repeat items appeared to be particularly challenging to score automatically, potentially because of short response duration.

In this research report, we further explore differences between item types when using automated speech scoring technology for non-native adult speech in the context of spoken English proficiency assessment. All of the main components of the *SpeechRater*^{SM} automated scoring engine (Higgins, Xi, Zechner, & Williamson, 2011; Zechner et al., 2009) from the *Educational Testing Service* (ETS) were adapted and modified to score 32 items encompassing 7 item types selected to target the evaluation of non-native speakers of English with low speaking proficiency (corresponding to A1–B2 bands within the Common European Framework of Reference for Languages [CEFR]). Language models (LMs) by the ASR system were adapted for each item type; filtering models to identify several categories of nonscorable responses were developed; and scoring models were built for each item type using a hybrid approach involving automated feature selection with a shrinkage method and expert review.

Table 1 lists the seven item types investigated in this study along with the number of items per item type, response complexity (i.e., entropy level), and the allocated response time for the test taker. In this report, we distinguish between low- and medium-entropy items. For low-entropy items, the correct response is entirely predictable. These are read-aloud (RA) and sentence-repeat (SR) items, which require the test taker to either read aloud the printed prompt or to repeat the recorded prompt. For medium-entropy items, we may expect the test taker to use particular words or phrases, but we cannot predict the test taker's response in advance. Specifically, for the medium-entropy agenda items (AG12 [agenda item/short] and AG3 [agenda item/long]), the test takers were asked to read a printed agenda and answer several questions based on this agenda. The first two questions (both AG12) required a relatively short response, whereas the last question (AG3) was more open ended and prompted the test taker to provide a longer and more elaborate response. The market survey (MS) item asked test takers to respond to a question (e.g., what considerations are important when buying a particular product?). Finally, for the picture description (PD) item, test takers were asked to give a detailed description of a picture. Note that the responses to the last two item types (MS and PD) are likely to be less predictable than the responses to agenda items, for which test takers are provided with printed materials. Therefore, these two items are classified as medium–high entropy.

Table 1 Overview of Item Types Used

Entropy level	Item type	Item type label	Total *N* of items	Allowed response duration per item (s)
Low	SR	Sentence repeat	8	10
Low	RA	Read aloud	4	30
Medium	AG12	Agenda/short	4	15
Medium	AG3	Agenda/long	4	30
Medium–high	PD	Picture description	4	45
Medium–high	MS	Market survey	4	30

The remaining sections of this report are organized as follows: we first describe the data used for this study; we then proceed to describe the language model adaptation work and related results, the details on the filtering models we developed for identifying nonscorable responses, and the item type-specific scoring models. The report concludes with the presentation of the performance of the whole system and a discussion of the findings.

Data

The data used in this study were drawn from several administrations of a global language proficiency assessment of English for workforce purposes, which included the preceding item types.

The main corpus used to evaluate and compare the model performance for different item types (referred to as the *evaluation set*) consisted of responses from 2,090 speakers collected as part of a pilot study. There were two test forms, with each form consisting of four SR, two RA, two PD, two MS, two AG12, and two AG3 items. Each speaker responded to one of the two forms. The speakers came from 13 different countries in Asia and South America, representing a variety of native languages.

The training data for different components of the automated scoring engine were drawn from the responses to the same items collected during previously administered operational and pilot language assessments. Such a combination of training and evaluation data mimics a hypothetical scenario where an existing database of responses is used to train the models for a new assessment consisting of similar items. However, it also leads to a mismatch between training and evaluation data because the two datasets are collected under different conditions. We will return to this issue later in the report. For the same reason, because the size of the training dataset is limited by the availability of existing data, contrary to what is common in many machine learning studies, the size of the evaluation set exceeds the size of the training set for some of the item types.

All responses in the training data were transcribed by professional transcribers and divided into several different subsets for system development, including training, development, and evaluation sets to tune the ASR as well as the training and development sets for building the filtering and scoring models. Table 2 lists the number of responses in different partitions across different item types. Note that the training data included in this table only include the responses to the items included in the evaluation set. For the ASR engine and filtering models, these datasets were supplemented with responses drawn from different assessments, as explained in the corresponding sections.

All responses were scored by at least one human rater, with every response in the evaluation set scored by two raters. Medium- and medium–high-entropy items (AG12, AG3, PD, and MS) were scored on a scale of integers 0–3. Low-entropy items (RA and SR) were assigned two scores, one for pronunciation and another for intonation, both on a scale of integers 0–3. In this report, the two scores were summed to obtain a single rating for each low-entropy item.

Automatic Speech Recognition

The first step in automated scoring is the transcription of spoken responses. The ASR system located at the front end of SpeechRater is an important component that impacts the performance of the whole system.

Table 2 Total Number of Responses in Different Partitions Used to Train and Evaluate the Automated Scoring System

	Training set adaptation data for ASR			Training set training data for scoring and filtering models		
Item type	Training	Development	Evaluation	Training	Development	Evaluation set
RA	1,200	200	200	1,200	1,200	4,180
SR	2,462	200	200	400	200	8,360
AG12	2,400	400	400	2,400	2,400	4,180
AG3	1,200	200	200	1,200	1,200	4,180
PD	1,200	200	200	1,200	1,200	4,179
MS	1,200	200	200	1,200	1,200	4,180

Note. Both training and evaluation sets contain responses to the same set of items. AG12 = agenda/short; AG3 = agenda/long; ASR = automatic speech recognition; MS = market survey; PD = picture description; RA = read aloud; SR = sentence repeat.

Table 3 Optimal Parameters for the Language Model (Lambda) and Decoder (Rejection Threshold) for Different Item Types and Adaptation Data

Adaptation	SR	RA	AG12	AG3	PD	MS
Prompt	.9, −25	.9, −25	n/a	n/a	n/a	n/a
Hypothesis	.9, −25	.8, 0	.8, −25	.8, −15	.8, −30	.7, −25
Transcript	.9, −25	.9, −25	.9, −25	.8, −15	.9, −25	.8, −15

Note. AG12 = agenda/short; AG3 = agenda/long; MS = market survey; PD = picture description; RA = read aloud; SR = sentence repeat.

An ASR system generally includes three configurable sources of information: the acoustic model, the language model, and the lexicon. The acoustic model (AM) is a statistical representation of knowledge about acoustics, phonetics, gender, speaker dialect differences, and so on. The LM incorporates knowledge of possible word sequence, semantics, and grammatical variation. The dictionary maps pronunciation units, such as phones from which the AM is constructed, to the words presented in the LM.

The speaker-independent AM for this study was trained using an external dataset not included in Table 2. This dataset contained approximately 800 hours of unscripted (spontaneous) responses to several administrations of an international language assessment, which used different types of items than the ones considered in this study. These responses were collected from non-native English test takers with diverse L1s and, therefore, can be seen as a generic database of non-native English speech.

To overcome the limited amount of data available for training the LMs and to obtain the best ASR performance, the language models were initially trained using the human transcriptions of the same generic data used for training the AMs. These LMs were then adapted to each item type considered in this study using three different sets of adaptation data: item prompt (low-entropy items only), transcribed responses to each item, and ASR hypotheses for these responses (the "training" partition of the "adaptation data for ASR" in Table 2). The ASR hypotheses used for adaptation were generated using the AM and LM trained on the original 800 hours of generic non-native speech. These different setups allowed us to explore several operational scenarios. Although in an ideal situation, human transcriptions may be available to fine-tune the ASR to new data, financial or other considerations may make collecting such transcriptions impractical. In this case, item prompts and/or ASR transcription of the responses might be used as a substitute for human transcriptions.

To further optimize ASR performance, two parameters, LM interpolation weight (lambda) and the decoder word rejection weight (rejection), were tuned using the ASR development partition. The lambda parameter determines the relative effect between broader contexts and more domain-specific contexts in LM, whereas the rejection weight sets the threshold for rejecting or accepting the ASR hypothesis for each word. The ranges of these two parameters varied between .6 and .9, with an increment of .1 for lambda and between −30 and 0 with an increment of 5 for the rejection threshold. The optimal parameter values for all item types are shown in Table 3.

The ASR performance was measured as WER on the ASR evaluation dataset. The WER for each item type using the optimal parameters are shown in Table 4. The model without any adaptation provides the lower bound for the ASR performance with the out-of-domain language model.

As can be seen from Table 4, all adaptation strategies lead to a substantial reduction in WER in comparison to the generic model. The adaptation based on human transcriptions leads to the lowest WER, and therefore, this method and its respective parameters (listed in Table 3) were used in the final ASR configuration. The hypothesis produced by the

Table 4 Word Error Rate Across Different Item Types

Adaptation	SR	RA	AG12	AG3	PD	MS
No adaptation	80.3	52.1	66.2	63.9	51.3	44.4
Prompt	38.9	10.4	n/a	n/a	n/a	n/a
Hypothesis	39.0	11.8	54.6	49.0	41.4	39.2
Transcript	30.3	7.5	34.3	28.3	32.4	35.2

Note. Values are in percentages. AG12 = agenda/short; AG3 = agenda/long; MS = market survey; PD = picture description; RA = read aloud; SR = sentence repeat.

Table 5 Proportion of Technical Difficulty (TD) and 0 Responses in the Entire Data (Rater 1)

Item type	TD responses (%)	0 responses (%)
SR	3.0	13.3
RA	4.2	3.1
AG12	3.7	12.6
AG3	2.9	10.9
PD	3.3	4.7
MS	4.0	5.8
Average	3.45	8.70

Note. AG12 = agenda/short; AG3 = agenda/long; MS = market survey; PD = picture description; RA = read aloud; SR = sentence repeat.

ASR system was then used along with the audio signal to compute 107 SpeechRater features covering various aspects of language proficiency as well as audio quality of the responses in this study. These features were then used to train the filtering and scoring models described in the following sections.

Filtering Models

Some spoken responses in English proficiency assessments tend to have suboptimal characteristics, which present serious challenges for automated speech scoring. In general, these problematic responses can be classified into two groups: responses with technical difficulty (hereinafter TD responses) and responses with a human score of 0 (hereinafter 0 responses). Typical TD responses are responses with severe audio problems that obscure the content of the responses and/or distort the spectral characteristics (e.g., responses with high noise levels). Responses in the 0 score group contain suboptimal characteristics from uncooperative test takers. Typical response types in this group include empty responses (no-speech response), responses in speakers' native languages (non-English response), and responses not related to the content of the prompt (off-topic response). Table 5 summarizes the proportion of TD and 0 responses in our data as assigned by the first human rater.

As can be seen from the table, the proportion of TD responses remained relatively consistent across different item types and averaged at 3.45%. On the contrary, the proportion of score 0 responses varied largely across item types, ranging from 3.1% to 13.3%, with more than 10% of the SR, AG12, and AG3 responses receiving a 0 score.

The automated scores for these problematic responses are likely to be erroneous, and the inclusion of these responses in the automated scoring model building process could reduce the validity of the automated scores (Zhang, Chen, & Ruan, 2015). To address this problem, SpeechRater uses a filtering model as a prescreening mechanism to identify problematic responses. The filtering models developed in this study classify responses into three classes: TD, 0, and scorable. The idea is that under the operational conditions where individual item scores are aggregated to provide a total score for each speaker, the responses classified as belonging to the TD group can be automatically filtered out and excluded from the speaker-level score aggregation. Responses classified as the 0 group would be assigned a score of 0 and can be used for the speaker-level score aggregation. Finally, the responses considered scorable are assigned a numeric score based on the scoring model and can be used for the speaker-level score aggregation.

Filtering models in this study were developed based on a subset of SpeechRater features. A total of 45 features were initially selected, which can be grouped into "basic features," "acoustic features," and "ASR features." These features were designed to monitor the audio recording and speech recognition-based automated transcription generation processes:

- *Basic features* (22 features): features related to basic information about the spoken response (e.g., the number of words, duration of speech segment, and frequency and duration of pauses based on speech recognition output)
- *Acoustic features* (16 features): features (e.g., the mean and standard deviation of energy and pitch) related to energy, pitch, and spectral characteristics of the audio signal
- *ASR features* (seven features): features derived from the automated speech recognition system to monitor the performance of the speech recognizer (e.g., normalized confidence score that is a self-diagnostic score of the speech recognizer's performance)

Table 6 Performance of Filtering Model for Each Item Type

Item type	Baseline accuracy (%)	Accuracy (%)	Precision	Recall	*f*-Score
SR	84	93	.74	.76	.75
RA	93	96	.71	.71	.71
AG12	84	93	.75	.78	.77
AG3	86	94	.74	.75	.74
PD	92	96	.73	.76	.75
MS	90	95	.73	.76	.74
Total	88	94	.75	.77	.76

Note. Precision, recall, and *f*-scores were averaged over the three categories TD, 0, and scorable. AG12 = agenda/short; AG3 = agenda/long; MS = market survey; PD = picture description; RA = read aloud; SR = sentence repeat.

These three groups of features were used for a three-class classification using a decision tree model based on the J48 algorithm (WEKA [Frank, Hall, & Witten, 2016] implementation of C4.5). One generic model was trained using a combined sample from different item types. To overcome the data sparseness issue due to the low proportion of TD responses, we extracted 800 TD responses from the operational administrations of a different international language assessment and added these to the model training partition described in Table 2 to bring the total share of TD responses to 5%.

The filtering system was evaluated using the scoring model evaluation partition. Table 6 provides the performance of the filtering model for each item type.

The accuracy of the filtering system was 94%, and there was a relative 50% error reduction (from 12% to 6%) compared to a simple majority class-based system where all responses were classified as scorable responses. The accuracy of the filtering model was comparable across the six item types, ranging from 93% to 96%.

In addition to the accuracy, we calculated the average precision, recall, and *f*-score across the three classes. The system overall achieved a good performance; the *f*-scores of all item types were over .70, and the average was .76. Compared to the accuracy, there were larger variations across different item types, particularly for recall and *f*-score. The item type with the lowest *f*-score was RA (.71). The low performance of the filtering model for RA items might have resulted from the sparseness of the 0 responses; that is, the proportion of 0 responses for this item type was substantially lower than other item types.

Scoring Models

The final step in training the automated scoring system is to build a scoring model for scorable responses that maps the feature values to the final scores. In this study, separate scoring models were built for each item type to maximize the construct coverage and model performance for each item type.

Following the standard procedure for building scoring models, the feature values were first transformed to ensure that their distribution was close to normal and were standardized using *z*-scores. For features wherein the raw values were expected to be negatively correlated with human scores, the values were multiplied by −1 to ensure that all expected human–feature correlations are positive. Finally, the scoring models were built using the subset of responses in the training set that received numeric nonzero human scores and had no missing feature values.

For each item type, we used a hybrid approach described in Loukina, Zechner, Chen, and Heilman (2015) to identify an optimal feature set that would satisfy certain requirements of the scoring model. In this approach, an expert in automated speech scoring first identifies a subset of features that might be appropriate for a particular item type. As the next step, penalized Lasso regression is used to select the subset of features that yields an optimal balance between the model performance and the total number of features in the model. Finally, an expert in speaking evaluation reviews the model and makes adjustments, if necessary, to ensure proper construct coverage.

The resulting scoring models consisted of 8–16 features, depending on the item type. The coefficients for each feature were estimated using ordinary least squares linear regression on the training set.

The following sections show the construct coverage and performance for each of the six item types discussed in this report.

Table 7 Construct Coverage of Automated Scoring Models for Different Item Types (Number of Features Representing Each Construct)

	Medium – high entropy		Medium entropy		Low entropy	
Construct coverage	MS	PD	AG12	AG3	SR	RA
Delivery	11	8	9	9	6	11
Language use	2	3	2	6	N/A	N/A
Content accuracy	1	1	1	1	2	2
Total *N* features	14	12	12	16	8	13

Note. AG12 = agenda/short; AG3 = agenda/long; MS = market survey; PD = picture description; RA = read aloud; SR = sentence repeat.

Model Construct Coverage

The content coverage of the scoring models was different for medium- and low-entropy items. Table 7 summarizes the construct coverage for each item type. All SpeechRater features used in scoring models are shown in Appendix A.

Low-Entropy Items

For low-entropy items (RA and SR), the scoring models included features representing two constructs: delivery and content accuracy. Delivery features different evaluated aspects of pronunciation and fluency, such as the number of disfluencies, overall pronunciation accuracy, stress patterns, and intonation. Because for both SR and RA items, the correct response is expected to fully match the prompt, the two content accuracy features, cwpm and lowentwer, measured how much the submitted response deviated from the prompt.

Medium- and Medium–High-Entropy Items

For medium- and medium – high-entropy items (PD, MS, AG12, and AG3), the scoring models covered three main constructs: delivery, language use, and content accuracy. Of these three constructs, the features representing the delivery construct were common to both low- and medium-entropy items. Although the models for low-entropy features also covered content accuracy, the approach to evaluating this construct was different for medium-entropy items. Because, by definition, no single correct reference response could be used to evaluate the content coverage, the test responses were instead compared to a large corpus of reference responses to each item using BLEU score (Papineni, Roukos, Ward, & Zhu, 2002; Zechner & Wang, 2013). This metric evaluates how many words in each response also occurred in the reference responses. Finally, language use features covered aspects of language proficiency, like grammar and vocabulary.

Scoring Model Performance

The scoring model was evaluated on the responses from the evaluation set that received numeric nonzero human scores and SpeechRater scores. We compared the interhuman agreements with the human – machine agreements for each item type. The results are given in Tables 8 and 9. Of note is that the results in the tables are aggregated results across items of the same item type. Model performance on individual items is provided in Appendix B for interhuman agreement and Appendix C for the corresponding machine – human agreement.

In general, human – human agreement was comparable to what has been reported for other assessments of spoken language proficiency. For example, in Loukina et al. (2015), human – human agreement for high-entropy items was .62 at the item level, whereas in Zechner et al. (2014), the human – human agreement varied between .51 and .83 for low-entropy items and between .67 and .80 for medium-entropy items. In this study, we found that the interhuman agreement was higher for the AG12, AG3, and SR items, with correlation coefficients and quadratically weighted kappa values greater than .70. The human raters also achieved a reasonable level of exact agreement (i.e., exact percentage agreement was 76.7% and 80.8% for AG12 and AG3, respectively, on a score scale of 1 – 3 and 50.5% for SR on a score scale of 1 – 6). In contrast, human raters did not agree with one another well for grading responses in the remaining three item types, particularly PD. The human – human correlation coefficient and quadratically weighted kappa were only .60 for PD, and the exact percentage agreement was only 66.5%, suggesting that more than 30% of pairs of human ratings differed by 2 points on

Table 8 Human – Human Agreement (Nonzero Scores Only)

				Score distribution				Human1 – Human2 agreements				
Item type	Score scale	Item counts	Sample size	Mean (H1)	*SD* (H1)	Mean (H2)	*SD* (H2)	*R*	QWK	% Agree	Adj % Agree	SMD
MS	1–3	4	3,648	2.19	.63	2.19	.63	.64	.64	72.75	99.67	.00
PD	1–3	4	3,772	2.08	.66	2.10	.65	.60	.60	66.54	99.52	.03
AG12	1–3	4	3,300	1.85	.73	1.86	.74	.78	.78	76.70	99.73	.02
AG3	1–3	4	3,470	1.96	.61	1.97	.62	.74	.74	80.78	99.97	.02
RA	2–6	4	3,754	4.49	1.05	4.54	1.06	.65	.65	53.89	89.58	.05
SR	2–6	8	6,587	3.84	1.39	3.82	1.37	.72	.72	50.54	83.33	−.01

Note. Adj % Agree = 1-point adjacent percentage agreement; % Agree = exact percentage agreement; AG12 = agenda/short; AG3 = agenda/long; H1 and H2 = human raters; K = unweighted kappa; MS = market survey; PD = picture description; QWK = quadratically weighted kappa; *R* = Pearson correlation coefficient; RA = read aloud; SMD = standardized mean score difference; SR = sentence repeat.

Table 9 Human – SpeechRater Agreement (Numeric, Nonzero Scores Only)

				Score distribution				Human1 – Human2 agreements				
Item type	Score scale	Item counts	Sample size	Mean (H1)	*SD* (H1)	Mean (SpR)	*SD* (SpR)	*R*	QWK	%Agree	Adj %Agree	SMD
MS	1–3	4	3,668	2.19	.63	2.24	.38	.49	.39	61.80	99.07	.10
PD	1–3	4	3,791	2.07	.66	2.20	.45	.63	.53	64.20	99.63	.22
AG12	1–3	4	3,321	1.84	.73	1.93	.53	.59	.50	56.76	98.92	.14
AG3	1–3	4	3,492	1.96	.61	1.98	.52	.59	.50	65.29	99.71	.03
RA	2–6	4	3,766	4.49	1.05	4.70	.74	.66	.58	44.98	94.11	.23
SR	2–6	8	6,632	3.83	1.39	3.70	.94	.67	.60	33.62	86.22	−.11

Note. Adj %Agree = 1-point adjacent percentage agreement; %Agree = exact percentage agreement; AG12 = agenda/short; AG3 = agenda/long; H1 and H2 = human raters; K = unweighted kappa; MS = market survey; PD = picture description; QWK = quadratically weighted kappa; *R* = Pearson correlation coefficient; RA = read aloud; SMD = standardized mean score difference; SpR = SpeechRater; SR = sentence repeat.

a 1–3 scale. On the basis of this sample, these item types with low human – human agreements were easier items for test takers. For example, the average Human1 score for MS items was 2.19 on the 3-point scale, which might have led to a ceiling effect incurring low interhuman agreements.

The automated scoring models appeared to perform somewhat better for low-entropy items — that is, the RA and SR items — as well as for the medium-entropy PD item. For these items, the correlations between the predicted and observed human scores were comparable to the correlations between two human raters. However, for some low-entropy items, the automated scoring models appeared to have introduced potential biases in the scoring process. For example, the automated scoring models awarded higher scores than human raters did in scoring the PD and RA items. The standardized mean score difference (SMD) values for these two items were .22 and .23, respectively, considerably larger than the conventionally acceptable range for this index (i.e., .15; Williamson, Xi, & Breyer, 2012). For the other four item types, the SMD values were within the recommended ranges. We also note that the model performance varied across individual items for all item types (shown in Appendix D). The agreement was generally lower for medium-entropy items, with a particularly low performance for MS. For this item type, the correlation between the predicted and observed scores was consistently below .5 for all prompts.

Performance of the Complete Automated Scoring System

Finally, we evaluated the performance of the complete SpeechRater scoring system, which combined both the automatic filtering model and the automatic scoring model. For this evaluation, all responses that were classified as TD or 0 by the filtering model were automatically assigned TD or 0 correspondingly. In addition to responses classified as TD by the filtering model, there were also responses that were classified as scorable but for which some of the SpeechRater features

Table 10 Overview of the Combined Filtering and Scoring Model

Filtering model output	Scoring model output	Final score
Scorable	Numeric score	Numeric nonzero score as computed by the scoring model
	No score produced due to missing feature values	TD
0	–	0
TD	–	TD

Note. TD = technical difficulty.

Table 11 Percentage of Responses Assigned Technical Difficulty (TD), 0, or Numeric Scores by the Final System

	TD			
Item type	Based on FM output	Due to non-numeric feature values	0	Numeric score
RA	3.8	2.2	2.7	91.3
SR	4.4	3.5	11.0	81.1
AG12	5.4	2.7	11.1	80.8
AG3	4.7	1.1	9.0	85.2
PD	4.3	.0	3.6	92.1
MS	4.5	.9	5.7	88.9

Note. See Table 5 for values for human raters. Values are in percentages. AG12 = agenda/short; AG3 = agenda/long; FM = filtering model; MS = market survey; PD = picture description; RA = read aloud; SR = sentence repeat.

could not be computed. These missing feature cases were primarily due to either a very short ASR hypothesis, which made it impossible to compute such features as a relative percentage of stressed syllables, or low recording quality, which resulted in pitch extraction failure. For these kinds of responses, the automated scoring model can only use a subset of the full feature set, and therefore, the resulting automated scores are not valid. Thus, we classified them as TD as well. We only used the scoring model to compute scores for all responses that have been classified as scorable by the filtering model and had no missing feature values. The procedure used to combine scoring and filtering models is summarized in Table 10.

Table 11 shows the distribution of the responses in each category for different item types. Although the percentage of responses classified as TD was similar across different item types, there was a notable variation in the percentage of responses where the feature extraction could not be completed. This number was particularly high for SR items where 3.5% of all responses had non-numeric features.

We then performed the final evaluation using all responses that received numeric scores (including zero scores) from both SpeechRater and the first human rater. It should be noted that the inclusion of responses with zero scores in our analyses naturally led to an increment of the score scale for each item along with a slightly greater sample size. The aggregated results on the item type level are shown in Tables 12 and 13, and the corresponding item level results are given in Appendices D and E.

The results are rather similar to what we found using only responses with numeric nonzero scores. That is, the human–machine agreement was the highest for the two low-entropy items and PD item.

Discussion

In this report, we have reviewed the performance of different components of an automatic speech scoring system for different low- to medium-entropy item types. The performance of different components of the system is summarized in Table 14.

The table shows that, unsurprisingly, the complete automated scoring system generally performed better for low-entropy items and best for RA items, which had the lowest ASR WER. The correlation between system scores and human scores for these items was the same as for two human raters. In agreement with what has been reported in previous studies (Cheng et al., 2014; Zechner et al., 2014), the performance was worse for the other low-entropy item, SR, which yielded a higher WER and a small degradation between system performance versus two human raters. As in previous studies, in this corpus, SR items were also the shortest items—which may partially explain the low system performance.

Table 12 Human–Human Agreements (Evaluation Set Including Human Score of Zero)

				Score distribution				Human1–Human2 agreements				
Item type	Score scale	Item counts	Sample size	Mean (H1)	*SD* (H1)	Mean (H2)	*SD* (H2)	*r*	QWK	%Agree	Adj %Agree	SMD
MS	0–3	4	3,679	2.17	.65	2.17	.65	.66	.66	72.66	99.59	.00
PD	0–3	4	3,807	2.06	.69	2.08	.67	.62	.62	66.51	99.47	.03
AG12	0–3	4	3,357	1.82	.76	1.83	.76	.79	.79	76.29	99.64	.02
AG3	0–3	4	3,538	1.93	.65	1.94	.66	.76	.76	80.36	99.69	.02
RA	0–6	4	3,776	4.46	1.10	4.52	1.10	.66	.66	53.87	89.35	.05
SR	0–6	8	6,738	3.76	1.47	3.74	1.46	.74	.74	50.74	82.80	−.01

Note Adj%Agree = 1-point adjacent percentage agreement; %Agree = exact percentage agreement; AG12 = agenda/short; AG3 = agenda/long; H1 and H2 = human raters; K = unweighted kappa; MS = market survey; PD = picture description; QWK = quadratically weighted kappa; *r* = Pearson correlation coefficient; RA = read aloud; SMD = standardized mean score difference; SR = sentence repeat.

Table 13 Human–SpeechRater Agreements (Evaluation Set, Combined Scores From Filtering and Scoring Models)

				Score distribution				Human1–SpeechRater agreements				
Item type	Score scale	Item counts	Sample size	Mean (H1)	*SD* (H1)	Mean (SpR)	*SD* (SpR)	*r*	QWK	%Agree	Adj %Agree	SMD
MS	0–3	4	3,691	2.17	.65	2.23	.39	.50	.40	61.42	98.81	.11
PD	0–3	4	3,822	2.06	.69	2.19	.46	.64	.53	63.68	99.27	.23
AG12	0–3	4	3,365	1.82	.76	1.92	.53	.60	.50	56.02	98.66	.16
AG3	0–3	4	3,544	1.93	.65	1.96	.53	.61	.52	64.33	99.55	.06
RA	0–6	4	3,785	4.46	1.10	4.69	.75	.66	.57	44.76	93.63	.24
SR	0–6	8	6,750	3.76	1.47	3.68	.95	.67	.59	33.04	84.71	−.07

Note. Adj % Agree = 1-point adjacent percentage agreement; % Agree = exact percentage agreement; AG12 = agenda/short; AG3 = agenda/long; H1 and H2 = human raters; K = unweighted kappa; MS = market survey; PD = picture description; QWK = quadratically weighted kappa; *r* = Pearson correlation coefficient; RA = read aloud; SMD = standardized mean score difference; SpR = SpeechRater; SR = sentence repeat.

Table 14 Summary of Performance of Different Components of the Automated Scoring System for Different Item Types Considered in This Report

Item type	Entropy	Response duration	Total % of nonscorable responses	ASR performance (WER)	Filtering model performance (*f*-score)	Full system (*r*)	Human–human agreement (*r*)	Degradation
SR	Low	10	18.9	30.3	.75	.67	.74	.07
RA	Low	30	8.7	7.50	.71	.66	.66	0
AG12	Medium	15	19.2	34.3	.77	.60	.79	.19
AG3	Medium	30	14.8	38.3	.74	.61	.76	.15
PD	Medium–high	45	7.9	32.4	.75	.64	.62	−.02
MS	Medium–high	30	11.1	35.2	.74	.50	.66	.16

Note. AG12 = agenda/short; AG3 = agenda/long; MS = market survey; PD = picture description; RA = read aloud; SR = sentence repeat; WER = word error rate.

For the medium–high-entropy items, PD items showed the best performance, with the correlation between system scores and human scores exceeding the correlation between two human raters. For other medium-entropy items, the overall performance showed degradation in comparison to human scores. The moderate system performance was partially due to the mismatch between the training and evaluation data discussed earlier. Appendix F shows the system performance evaluated on a dataset that was collected together with the data used for training the model and matched these data in terms of test-taker population and experimental conditions. The correlation coefficient between the system scores and human scores approaches .7 for most item types, and the standardized mean differences are generally under the recommended threshold of .15. At the same time, we see similar patterns of performance for different item types. In both cases, the performance was the lowest for MS and AG12 items. There was a notable difference in AG3 items, which showed good performance when the model was trained on matching data but a somewhat lower performance for mismatching data. In general, our results once again show that matching the training data and the data expected from a new assessment by item type results in the moderate performance of the models. The performance can be improved by matching the data not only in terms of item type but also in terms of test-taker population and data collection method.

We also considered the performance of each component of the speech scoring engine. Although the performance of the scoring model across different item types closely mirrored the performance of the complete system, the same cannot be said about all system components. Thus, the item type had little effect on the performance of the filtering models wherein the f-score remained relatively constant. While ASR performance was generally better for low-entropy items, there also was a striking difference between the two types of such items—with the WER for SR being four times higher than the WER for RA. Therefore, when selecting the most suitable items for automated scoring, the performance of each component should be considered independently, taking into account the nature of the assessment and the role of the automated scoring system in determining the final score.

Another important consideration is the number of nonscorable responses to each item type. The percentage was generally the lowest for RA and PD items (7–8%), which were also the items associated with the highest performance of the scoring model. We found that the items with very short response times (10–15 s), such as SR and AG12, also had a very high percentage of nonscorable responses (18%)—including a higher percentage of responses with feature extraction failure (approximately 3%). This, along with low scoring model performance for at least some short-response items observed in both this and previous studies, suggests that short items should probably be avoided when designing assessments that are expected to be scored using an automated scoring engine.

Finally, although in this study, we used human transcriptions to adapt the LMs to these particular items, we found that in the absence of such transcriptions, ASR hypotheses of the responses in combination with prompt information (when available) may also be used to improve performance in comparison to the generic model.

Conclusion

In this report, we have compared the performance of different components of the automatic speech scoring engine on several item types. We found that the agreement between final system scores and human scores was the highest for low-entropy items, but low-entropy items with short response times also showed lower ASR performance and yielded a high percentage of nonscorable responses. A similar trend was observed for medium-entropy items with short response times. Finally, we showed that the effect of item type varies across the system components. Therefore, when selecting the items for automated scoring, it is important to consider how each individual system component affects the reliability and validity of the final score and which components should be given priority given the nature of a particular assessment.

References

Bernstein, J., Cheng, J., Suzuki, M., Ave, S. C., & Alto, P. (2010). Fluency and structural complexity as predictors of L2 oral proficiency. *Proceedings of Interspeech 2010* (pp. 1241–1244). Chiba, Japan: International Speech Communication Association.

Bernstein, J., De Jong, J., Pisoni, D. B., & Townshend, B. (2000). Two experiments on automatic scoring of spoken language proficiency. In P. Delcloque (Ed.), *Proceedings of InStil2000* (pp. 57–61). Dundee, UK: University of Abertay.

Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 722–731). Portland, OR: Association for Computational Linguistics.

Cheng, J., D'Antilio, Y. Z., Chen, X., & Bernstein, J. (2014). Automatic assessment of the speech of young English learners. *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 12–21). Baltimore, MD: Association of Computational Linguistics.

Cucchiarini, C., Strik, H., & Boves, L. (1997). Automatic evaluation of Dutch pronunciation by using speech recognition technology. *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings* (pp. 622–629). Santa Barbara, CA: IEEE.

Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). *Automated scoring for the TOEFL Junior® comprehensive writing and speaking test* (Research Report No. RR-15-09). Princeton, NJ: Educational Testing Service. http://dx.doi.org.10.1002/ets2.12052

Franco, H., Neumeyer, L., Digalakis, V., & Ronen, O. (2000). Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, *30*(2), 121–130.

Frank, E., Hall, M., Witten, I. (2016). *WEKA workbench online appendix for data mining: Practical machine learning tools and techniques* (4th ed.). Burlington, MA: Morgan Kaufmann.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, *25*(2), 282–306.

Loukina, A., Zechner, K., & Chen, L. (2014). Automatic evaluation of spoken summaries: the case of language assessment. *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 68–78). Baltimore, MD: Association for Computational Linguistics.

Loukina, A., Zechner, K., Chen, L., & Heilman, M. (2015). Feature selection for automated speech scoring. *Proceedings of the enth workshop on Innovative Use of NLP for Building Educational Applications* (pp. 12–19). Denver, CO: Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the ACL* (pp. 311–318). Philadelphia, PA: Association for Computational Linguistics.

Somasundaran, S., Lee, C. M., Chodorow, M., & Wang, X. (2015). Automated scoring of picture-based story narration. *Proceedings of the tenth workshop on Innovative Use of NLP for Building Educational Applications* (pp. 42–48). Denver, CO: Association for Computational Linguistics.

Strik, H., Van De Loo, J., Van Doremalen, J., & Cucchiarini, C. (2010, September). *Practicing syntax in spoken interaction: Automatic detection of syntactic errors in non-native utterances* (Paper O3-04). Paper presented at the Interspeech Satellite Workshop on Second Language Studies – Acquisition, Learning, Education and Technology, Tokyo, Japan. Retrieved from http://www.gavo.t.u-tokyo.ac.jp/L2WS2010/papers/L2WS2010_O4-03.pdf

Townshend, B., Bernstein, B., Todic, O., & Warren, E. (1998). Estimation of spoken language proficiency. *Proceedings of STiLL – Speech Technology in Language Learning* (pp. 93–96). Marholmen, Sweden: International Speech Communication Association.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.

Xie, S., Evanini, K., & Zechner, K. (2012). Exploring content features for automated speech scoring. *NAACL HLT '12 proceedings of the 2012 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 103–111). Montreal, Canada: Association for Computational Linguistics.

Yoon, S., Bhat, S., & Zechner, K. (2012). Vocabulary profile as a measure of vocabulary sophistication. *Proceedings of the seventh workshop on the Innovative Use of NLP for Building Educational Applications* (pp. 180–189). Montreal, Canada: Association for Computational Linguistics.

Zechner, K., Evanini, K., Yoon, S.-Y., Davis, L., Wang, X., Chen, L., … Leong, C. W. (2014). Automated scoring of spe aking items in an assessment for teachers of English as a foreign language. *Proceedings of the ninth workshop on Innovative Use of NLP for Building Educational Applications* (pp. 134–142). Baltimore, MD: Association for Computational Linguistics.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883–895.

Zechner, K., & Wang, X. (2013). Automated content scoring of spoken responses in an assessment for teachers of English. *Proceedings of the eighth workshop on Innovative Use of NLP for Building Educational Applications* (pp. 73–81). Atlanta, GA: Association for Computational Linguistics.

Zhang, M., Chen, J., & Ruan, C. (2015). Evaluating the detection of aberrant responses in automated essay scoring. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, W. C. Wang (Eds.), *Quantitative psychology research* (pp. 191–208). New York, NY: Springer.

Appendix A SpeechRater Features Used in Scoring Models for Each Item Type

			Medium entropy				Low entropy	
Feature	Description	Construct coverage	AG12	AG3	MS	PD	SR	RA
Dpsec	Number of disfluencies per second	Delivery/fluency					x	
clausecount	Number of clauses	Delivery/fluency	x					
Ipc	Number of interruption points per clause	Delivery/fluency	x			x		
Ipcount	Number of interruption points	Delivery/fluency						x
Ipw	Number of edit disfluencies per words	Delivery/fluency			x			
Numsil	Number of silences	Delivery/fluency					x	x
Silmean	Mean of silence duration	Delivery/fluency		x	x			
silmeandev	Average overall absolute differences between each silence duration and silmean	Delivery/fluency	x	x		x		
Silpwd	Number of silences per word	Delivery/fluency		x	x	x		
Tpsecutt	Number of types per second excluding initial and final pause	Delivery/fluency			x			
Types	Number of word types	Delivery/fluency	x	x	x	x		
Withinclausesilmean	Average duration of all within-clause silences	Delivery/fluency		x	x	x		
Wpsec	Speaking rate in words per second	Delivery/fluency	x		x			x
Wpsecutt	Number of words per second	Delivery/fluency		x			x	
Confavg	Average of word confidence scores	Delivery/pronunciation						x
conftimeavg	Confidence score per second	Delivery/pronunciation			x	x		x
L1	AM score of all the words	Delivery/pronunciation			x	x		
L3	AM score per phone	Delivery/pronunciation		x				x
L5	Average AM score density across all words	Delivery/pronunciation	x					
L6	Normalized acoustic model score	Delivery/pronunciation						x
L7	L5 normalized by the rate of speech	Delivery/pronunciation					x	
pitdeltanorm	Range of normalized pitch	Delivery/prosody	x				x	x
relstresspct	Relative frequency of stressed syllables in percentage	Delivery/prosody					x	x
stresyllmean	Average of distances between stressed syllables in syllables	Delivery/prosody	x	x				
tonesyllmean	Average of distances between syllables that bear tones in syllables	Delivery/prosody						x
Tonetimemdev	Mean deviation of distances between syllables that bear tones in seconds	Delivery/prosody			x			
Lmscore	Language Model score	Language use/grammar		x		x		
poscva4	CVA similarity between a response and responses with score 4 in grammatical expressions	Language use/grammar		x	x			
Avgfreq	The average frequency of word types in the response	Language use/vocabulary	x	x		x		
Avgrank	The average rank of word types in the response	Language use/vocabulary	x	x				
top2	The proportion of types that occurred both in a response and a reference list; reference list is TOP2 that is 101st–300th frequent word types inT2K-SWAL	Language use/vocabulary		x				

Appendix A continued

Feature	Description	Construct coverage	Medium entropy				Low entropy	
			AG12	AG3	MS	PD	SR	RA
top5	The proportion of types that occurred both in a response and a reference list; reference list is TOP5 that is 1501st–3,000th frequent word types in T2K-SWAL	Language use/vocabulary		x				
top6	The proportion of types that occurred both a response and a reference list; reference list is TOP6 that is over 3,001th frequent word types inT2K-SWAL	Language use/vocabulary	x					
Ttratio	Type-token-ratio	Language use/vocabulary		x		x		
bleu_s3	BLEU score by comparing the test response with sample responses from the highest score level	Content accuracy	x	x	x	x		
Cwpm	Correctly read words per minute	Content accuracy					x	x
lowentwer	WER between prompts and ASR hypotheses for each response using NIST's sclite scoring toolkit	Content accuracy					x	x
phn_shift	The mean of absolute shifts of the normalized vowel durations compared to standard normalized vowel durations estimated on a native speech corpus	Delivery/pronunciation			x			x

Appendix B Human–Human Agreement by Item (Numeric, Nonzero Scores Only)

				Score distribution				Human1–Human2 agreements				
Item type	Item ID	Score scale	Sample size	Mean (H1)	*SD* (H1)	Mean (H2)	*SD* (H2)	*r*	QWK	%Agree	Adj%Agree	SMD
AG12	VC456174	1–3	810	2.07	.80	2.10	.80	.79	.79	75.93	99.14	.04
	VC456175	1–3	777	1.62	.67	1.63	.69	.73	.73	76.19	99.74	.01
	VE030384	1–3	848	1.99	.65	2.01	.63	.70	.70	75.59	100.00	.03
	VE030385	1–3	865	1.69	.71	1.70	.71	.79	.79	78.96	100.00	.01
AG3	VC456176	1–3	772	1.60	.62	1.64	.62	.66	.66	73.96	100.00	.07
	VE029136	1–3	881	1.93	.59	1.92	.60	.73	.73	81.38	99.89	−.01
	VE030386	1–3	899	2.07	.50	2.07	.52	.68	.68	83.43	100.00	.01
	VE032709	1–3	918	2.19	.57	2.21	.58	.75	.75	83.33	100.00	.02
MS	VC360546	1–3	928	2.18	.54	2.24	.57	.58	.57	73.60	100.00	.10
	VC519218	1–3	901	2.14	.62	2.12	.62	.66	.66	74.69	99.67	−.03
	VE041817	1–3	874	2.06	.71	2.05	.69	.71	.71	72.43	99.66	−.01
	VE041833	1–3	945	2.37	.59	2.33	.59	.55	.55	70.37	99.37	−.07
PD	VC473433	1–3	941	2.06	.67	2.09	.64	.61	.61	67.38	99.57	.04
	VC517094	1–3	936	2.11	.68	2.09	.65	.64	.64	68.59	99.79	−.04
	VC976655	1–3	952	2.04	.66	2.06	.67	.59	.59	65.13	99.47	.02
	VC976778	1–3	943	2.08	.64	2.15	.64	.55	.55	65.11	99.26	.10
RA	VC382222	2–6	931	4.43	1.03	4.50	1.03	.65	.65	54.56	90.55	.07
	VC382238	2–6	922	4.55	1.03	4.68	1.08	.62	.61	50.76	88.29	.12
	VE033264	2–6	950	4.62	1.05	4.55	1.02	.65	.65	55.05	90.11	−.07
	VE033303	2–6	951	4.36	1.08	4.43	1.12	.68	.68	55.10	89.38	.07
SR	VE312049	2–6	791	3.54	1.32	3.61	1.33	.66	.66	47.41	80.40	.05
	VE312050	2–6	843	3.78	1.31	3.87	1.32	.67	.67	49.70	81.61	.07
	VE312051	2–6	883	3.66	1.24	3.75	1.29	.60	.60	46.66	79.28	.07
	VE312052	2–6	896	4.27	1.28	4.12	1.26	.66	.66	49.00	83.37	−.12
	VE312081	2–6	825	4.86	1.29	4.84	1.29	.74	.73	56.00	86.91	−.01
	VE312082	2–6	764	3.77	1.27	3.65	1.29	.72	.72	53.80	87.04	−.10
	VE312083	2–6	777	3.24	1.28	3.17	1.19	.67	.67	50.84	85.33	−.06
	VE312084	2–6	808	3.47	1.45	3.48	1.38	.74	.74	51.49	83.29	.01

Appendix C Human–SpeechRater Agreements by Item (Numeric, Nonzero Scores Only)

				Score distribution				Human1–SpeechRater agreements				
Item type	Item ID	Score scale	Sample size	Mean (H1)	*SD* (H1)	Mean (SpR)	*SD* (SpR)	*r*	QWK	%Agree	Adj %Agree	SMD
AG12	VC456174	1–3	816	2.07	.80	2.17	.54	.63	.52	50.86	98.77	.15
	VC456175	1–3	785	1.62	.67	1.73	.48	.56	.45	60.00	99.49	.19
	VE030384	1–3	851	1.99	.65	1.92	.49	.59	.47	61.69	99.65	−.12
	VE030385	1–3	869	1.69	.71	1.90	.51	.50	.41	54.55	97.81	.33
AG3	VC456176	1–3	782	1.59	.62	1.70	.52	.56	.46	63.04	99.74	.19
	VE029136	1–3	883	1.92	.59	1.97	.50	.53	.45	67.95	99.21	.08
	VE030386	1–3	905	2.07	.51	2.13	.53	.53	.43	61.99	100.00	.13
	VE032709	1–3	922	2.19	.57	2.06	.43	.59	.46	67.90	99.89	−.26
MS	VC360546	1–3	928	2.18	.54	2.27	.38	.48	.41	67.13	99.68	.19
	VC519218	1–3	906	2.14	.62	2.25	.37	.47	.40	63.80	98.79	.22
	VE041817	1–3	887	2.04	.72	2.16	.42	.52	.36	53.55	98.53	.19
	VE041833	1–3	947	2.37	.60	2.27	.35	.45	.35	62.41	99.26	−.20
PD	VC473433	1–3	943	2.06	.67	2.19	.45	.64	.52	63.41	99.68	.21
	VC517094	1–3	942	2.11	.68	2.18	.47	.64	.56	66.45	99.36	.12
	VC976655	1–3	956	2.04	.66	2.17	.44	.63	.51	63.39	99.79	.24
	VC976778	1–3	950	2.08	.64	2.25	.44	.61	.51	63.58	99.68	.32
RA	VC382222	2–6	937	4.42	1.04	4.70	.74	.67	.59	46.10	93.81	.30
	VC382238	2–6	926	4.55	1.03	4.68	.71	.63	.55	42.12	95.03	.15
	VE033264	2–6	950	4.62	1.05	4.88	.75	.64	.57	43.26	93.47	.29
	VE033303	2–6	953	4.36	1.08	4.52	.73	.69	.60	48.37	94.12	.18

Appendix C Continued

				Score distribution				Human1 – SpeechRater agreements				
Item type	Item ID	Score scale	Sample size	Mean (H1)	*SD* (H1)	Mean (SpR)	*SD* (SpR)	*r*	QWK	%Agree	Adj %Agree	SMD
SR	VE312049	2–6	797	3.53	1.32	3.44	.91	.63	.57	35.51	85.57	−.08
	VE312050	2–6	845	3.78	1.31	3.56	.81	.60	.51	33.73	83.79	−.20
	VE312051	2–6	885	3.66	1.24	3.56	.82	.60	.52	34.69	88.47	−.10
	VE312052	2–6	902	4.26	1.29	3.98	.89	.66	.57	35.92	88.58	−.26
	VE312081	2–6	834	4.85	1.30	4.26	1.07	.67	.56	28.42	85.01	−.49
	VE312082	2–6	771	3.77	1.27	3.51	.92	.65	.60	36.58	88.20	−.23
	VE312083	2–6	780	3.24	1.28	3.44	.78	.60	.50	30.90	85.90	.19
	VE312084	2–6	818	3.45	1.45	3.79	.99	.72	.62	33.13	83.99	.27

Appendix D Human–Human Agreements by Item (All Numeric Scores, Including Zero Scores)

				Score distribution				Human1 – Human2 agreements				
Item type	Item ID	Score scale	Sample size	Mean (H1)	*SD* (H1)	Mean (H2)	*SD* (H2)	*r*	QWK	%Agree	Adj %Agree	SMD
AG12	VC456174	0–3	829	2.03	.84	2.06	.84	.80	.80	75.51	98.91	.03
	VC456175	0–3	790	1.60	.69	1.61	.71	.74	.74	75.57	99.62	.01
	VE030384	0–3	861	1.97	.68	1.98	.67	.73	.73	75.38	100.00	.02
	VE030385	0–3	877	1.67	.73	1.68	.73	.80	.80	78.56	100.00	.01
AG3	VC456176	0–3	805	1.54	.67	1.58	.68	.71	.71	73.54	99.88	.05
	VE029136	0–3	893	1.90	.62	1.90	.64	.76	.76	81.41	99.78	.00
	VE030386	0–3	913	2.05	.54	2.05	.56	.70	.70	82.69	99.56	.00
	VE032709	0–3	927	2.18	.59	2.19	.61	.75	.74	82.96	99.57	.02
MS	VC360546	0–3	932	2.17	.56	2.23	.59	.60	.60	73.71	100.00	.10
	VC519218	0–3	909	2.13	.64	2.11	.64	.67	.67	74.48	99.45	−.03
	VE041817	0–3	887	2.03	.74	2.03	.73	.73	.73	72.04	99.55	−.01
	VE041833	0–3	951	2.35	.62	2.31	.61	.59	.59	70.45	99.37	−.07
PD	VC473433	0–3	950	2.05	.69	2.08	.67	.63	.63	67.37	99.47	.04
	VC517094	0–3	946	2.09	.71	2.07	.67	.67	.67	68.71	99.79	−.03
	VC976655	0–3	962	2.02	.69	2.04	.69	.61	.61	64.97	99.38	.03
	VC976778	0–3	949	2.07	.65	2.13	.66	.57	.57	65.02	99.26	.09
RA	VC382222	0–6	938	4.40	1.07	4.47	1.09	.66	.66	54.48	90.19	.06
	VC382238	0–6	924	4.54	1.05	4.67	1.10	.63	.63	5.87	88.31	.12
	VE033264	0–6	953	4.61	1.08	4.54	1.04	.65	.65	55.09	90.03	−.06
	VE033303	0–6	961	4.31	1.16	4.41	1.15	.69	.69	54.94	88.87	.08
SR	VE312049	0–6	821	3.43	1.43	3.49	1.44	.70	.70	47.87	79.66	.05
	VE312050	0–6	861	3.71	1.40	3.79	1.41	.71	.71	5.41	81.65	.06
	VE312051	0–6	890	3.64	1.27	3.72	1.33	.62	.62	46.85	79.21	.06
	VE312052	0–6	911	4.22	1.35	4.05	1.35	.69	.68	49.07	82.88	−.12
	VE312081	0–6	840	4.78	1.41	4.76	1.41	.76	.76	55.95	86.31	−.02
	VE312082	0–6	785	3.69	1.37	3.56	1.39	.75	.75	53.89	86.24	−.09
	VE312083	0–6	802	3.15	1.37	3.09	1.27	.71	.71	51.37	84.79	−.04
	VE312084	0–6	828	3.40	1.50	3.41	1.44	.75	.75	51.09	82.13	.00

Appendix E Human – SpeechRater Agreements by Item (All Numeric Scores, Including Zero Scores)

		Score	Sample	Score distribution				Human1 – SpeechRater agreements				
Item type	Item ID	scale	size	Mean (H1)	*SD* (H1)	Mean (SpR)	*SD* (SpR)	*r*	QWK	%Agree	Adj %Agree	SMD
AG12	VC456174	0 – 3	830	2.03	.84	2.15	.55	.65	.53	50.00	98.19	.17
	VC456175	0 – 3	795	1.60	.69	1.73	.48	.57	.45	59.25	99.12	.22
	VE030384	0 – 3	862	1.96	.68	1.91	.49	.60	.49	60.90	99.54	−.08
	VE030385	0 – 3	878	1.67	.73	1.89	.52	.51	.43	53.99	97.84	.34
AG3	VC456176	0 – 3	809	1.54	.68	1.68	.53	.59	.48	60.94	99.51	.23
	VE029136	0 – 3	894	1.90	.62	1.96	.51	.55	.46	67.11	98.88	.10
	VE030386	0 – 3	913	2.05	.54	2.12	.54	.56	.46	61.45	100.00	.14
	VE032709	0 – 3	928	2.18	.59	2.06	.45	.62	.47	67.46	99.78	−.24
MS	VC360546	0 – 3	932	2.17	.56	2.27	.38	.49	.42	66.85	99.46	.19
	VC519218	0 – 3	912	2.12	.64	2.25	.38	.50	.42	63.38	98.57	.23
	VE041817	0 – 3	895	2.02	.74	2.15	.42	.52	.36	53.07	97.99	.21
	VE041833	0 – 3	952	2.35	.62	2.26	.36	.48	.38	62.08	99.16	−.18
PD	VC473433	0 – 3	951	2.05	.69	2.18	.46	.65	.53	62.88	99.37	.22
	VC517094	0 – 3	952	2.09	.71	2.17	.47	.65	.56	65.76	98.84	.14
	VC976655	0 – 3	965	2.02	.69	2.17	.44	.63	.51	62.80	99.38	.25
	VC976778	0 – 3	954	2.07	.65	2.25	.44	.61	.51	63.31	99.48	.32
RA	VC382222	0 – 6	941	4.40	1.07	4.69	.74	.66	.57	45.91	93.41	.31
	VC382238	0 – 6	928	4.54	1.05	4.68	.71	.63	.55	42.03	94.83	.15
	VE033264	0 – 6	953	4.61	1.08	4.88	.76	.64	.56	43.13	93.18	.29
	VE033303	0 – 6	963	4.31	1.16	4.51	.74	.68	.58	47.87	93.15	.21
SR	VE312049	0 – 6	821	3.43	1.43	3.42	.92	.63	.56	34.47	83.07	−.01
	VE312050	0 – 6	861	3.71	1.40	3.54	.82	.61	.50	33.10	82.23	−.15
	VE312051	0 – 6	890	3.64	1.27	3.55	.82	.60	.52	34.49	87.98	−.08
	VE312052	0 – 6	912	4.22	1.35	3.97	.90	.67	.58	35.53	87.61	−.22
	VE312081	0 – 6	845	4.78	1.41	4.24	1.08	.67	.57	28.05	83.91	−.44
	VE312082	0 – 6	787	3.69	1.37	3.49	.92	.65	.59	35.83	86.40	−.17
	VE312083	0 – 6	803	3.14	1.37	3.43	.79	.59	.47	30.01	83.44	.25
	VE312084	0 – 6	831	3.40	1.50	3.77	.99	.72	.61	32.61	82.67	.29

Appendix F Human – SpeechRater Agreements by Item Type for Matching Training and Evaluation Data

	Score		Score distribution				Human1 – SR agreements			
Item type	scale	Sample size	Mean (H1)	*SD* (H1)	Mean (SpR)	*SD* (SpR)	*r*	QWK	%Agree	SMD
MS	0 – 3	396	2.56	.55	2.56	.38	.62	.56	75	.10
PD	0 – 3	397	2.63	.55	2.58	.43	.69	.66	82	−.06
AG12	0 – 3	774	2.20	.77	2.20	.52	.66	.57	60	.04
AG3	0 – 3	378	2.26	.68	2.27	.49	.71	.62	67	.09
RA	0 – 6	395	4.95	1.03	5.02	.70	.69	.59	47	.08
SR	0 – 6	200	4.40	1.41	4.18	.98	.67	.61	34	−.16

Suggested citation:

Loukina, A., Zechner, K., Yoon, S.-Y., Zhang, M., Tao, J., Wang, X., … Mulholland, M. (2017). *Performance of automated speech scoring on different low- to medium-entropy item types for low-proficiency English learners* (Research Report No. RR-17-12). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12139

Action Editor: Keelan Evanini

Reviewers: Lei Chen and Aliaksei Ivanou